Tijdschrift voor Medisch Onderwijs, september 2003 | Vol. 22, nr. 5, p. 229-234

Beoordelaarsbetrouwbaarheid (niet) meten met behulp van Cohens kappa?

J. Pols, H.E.P. Bosveld

Samenvatting

Inleiding: *Bij onderzoek naar beoordelaarsovereenstemming wordt frequent geadviseerd om Cohens kappa (kappa) te gebruiken als de schaal waarop de beoordeling wordt weergegeven van nominaal of ordinaal meetniveau is. Kappa is aantrekkelijk omdat hij corrigeert voor overeenstemming tussen beoordelaars op basis van toeval en dus de 'werkelijke' overeenstemming laat zien.*

Probleem: *In een onderzoek naar beoordelaarsovereenstemming tussen studenten die een consult beoordeelden, merkten wij op dat kappa lage waarden kan aannemen terwijl de ongecorrigeerde proportie overeenstemming hoog is (bijvoorbeeld: kappa -.03 en p .91).*

Literatuur: *Kappa blijkt een te negatief beeld te geven van de beoordelaarsovereenstemming in situaties waarbij het fenomeen zeer frequent of juist weinig frequent geobserveerd wordt. In de literatuur staat dit bekend als de prevalentieafhankelijkheid van kappa. In ons onderzoek deed deze situatie zich voor. Binnen het medisch onderwijs zal die situatie zich vaak voordoen bij vaardigheidsexamens. Kappa blijkt eveneens een te lage waarde aan te nemen onder omstandigheden waarin observatoren het structureel niet eens zijn over het al dan niet voorkomen van een fenomeen. Dit staat bekend als het bias-effect.*

Conclusie: *Cohens kappa is minder geschikt als maat om de beoordelaarsbetrouwbaarheid in uit te drukken, zeker bij fenomenen met een hoge of lage prevalentie. Is er geen alternatief voor kappa, dan moet ook melding gemaakt worden van de proportie positieve en negatieve overeenstemming. (Pols J, Bosveld HEP. Beoordelaarsbetrouwbaarheid (niet) meten met behulp van Cohens kappa? Tijdschrift voor Medisch Onderwijs 2003;22(5):229-234.)*

Inleiding

Er zijn veel situaties in het medisch onderwijs waarbij het relevant is om een oordeel te geven over de overeenstemming tussen observaties van één en dezelfde beoordelaar of van meerdere beoordelaars, de intra- en interbeoordelaarsbetrouwbaarheid. Het gaat daarbij om vragen zoals: in welke mate oordelen examinatoren steeds op dezelfde manier bij schriftelijke, mondelinge of vaardigheidstoetsen? En in welke mate stemmen oordelen van verschillende examinatoren overeen?

De aanleiding tot dit artikel is een onderzoek dat wij deden naar de beoordelaarsbetrouwbaarheid van een groep van 10 studenten die consulten observeerden.[1] Zij deden dat aan de hand van een beoordelingslijst waarop zij aan konden geven of een consultonderdeel wel of niet aan de orde kwam, een taak die vergelijkbaar is met die van examinatoren bij een vaardigheidstoets.

Bij de voorbereiding van het onderzoek zochten we in de literatuur naar een statistische methode waarmee de beoordelaarsbetrouwbaarheid in dit soort situaties wordt vastgelegd. Voor nominale (wel/niet, et cetera) en ordinale (reeksen van het type: zeer goed, goed, onvoldoende, et cetera) data blijkt daarvoor frequent gebruik gemaakt te worden van de kappacoëfficiënt die Cohen in 1960

beschreef (Cohens kappa, verder kappa genoemd).[2] Kappa is met name aantrekkelijk, omdat hij corrigeert voor de overeenstemming die op basis van toeval ontstaat en daardoor een beeld geeft van de mate waarin de observatoren werkelijk overeenstemmen.

In ons onderzoek waren we met name geïnteresseerd in de overeenstemming tussen onze observatoren voor elk item apart. Kappa is hiervoor niet geschikt, aangezien hiermee de overeenkomst tussen twee beoordelaars wordt berekend. En daarvoor zijn meerdere items nodig. Om een indruk te krijgen van de overeenstemming voor elk item apart is daarom ook de proportie overeenstemming berekend. Mede door deze opzet werd de aandacht gevestigd op een kenmerk van kappa dat ons aan het twijfelen heeft gebracht over de bruikbaarheid ervan.

In dit artikel geven we een toelichting op kappa en een deel van onze onderzoeksresultaten. In de bespreking gaan we nader in op onze twijfels en laten we zien wat daarover wordt geschreven in de internationale literatuur.

Cohens kappa

De algemene formule voor het corrigeren van een proportie overeenstemming voor toevalsovereenstemming is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Waarbij P_o = proportie geobserveerde overeenstemming (*P-observed*) en P_e = proportie overeenkomst op basis van toeval (*P-expected*).

Hoe groot de kans op toevallige overeenstemming (P_e) is, kan niet altijd op dezelfde manier worden vastgesteld. Wanneer bijvoorbeeld bij een multiple choice toets gecorrigeerd wordt voor het toevallig goed raden van een antwoord, hangt P_e af van het aantal alternatieven bij de vraag. Bij tweekeuzevragen is P_e dan .5 en bij vierkeuzevragen .25.

De eenvoudige kansberekening die gebruikt kan worden om P_e te berekenen bij meerkeuzevragen, is niet bruikbaar om vast te stellen hoe groot de kans is dat twee beoordelaars bij toeval dezelfde waarneming doen (bijvoorbeeld de kans dat twee examinatoren hetzelfde oordeel vellen over het verhelderen van de hoofdklacht tijdens een anamnese). Die kans moet geschat worden en dat gebeurt bij kappa op een specifieke manier. De werkwijze kan het best worden geïllustreerd aan de hand van een voorbeeld. Tabel 1 laat een 2 x 2 tabel zien met de resultaten van twee fictieve observatoren die elk een aantal keer aangaven of een gebeurtenis wel of niet plaatsvond. In 45% van de gevallen oordeelden beiden dat de gebeurtenis wel plaatsvond (P_{pos}), over 35% van de gevallen oordeelden zij beiden dat de gebeurtenis niet plaatsvond (P_{neg}). De totale proportie overeenstemming (P_{tot}) van de gebeurtenis vinden we op de diagonaal van de kruistabel: .45 + .35 = .80.

In 20% van de gevallen verschillen ze van mening over het al dan niet voorkomen van de gebeurtenis.

Tabel 1. Proporties waarnemingen van twee fictieve observatoren.

		Observator A		
		Wel	Niet	Totaal
Observator B	Wel	.45	.10	.55
	Niet	.10	.35	.45
	Totaal	.55	.45	1

Bij kappa wordt voor de berekening van P_e uitgegaan van de randtotalen. Er wordt aangenomen dat de beoordelaars onafhankelijk van elkaar zijn. In dat geval is de kans dat zij alleen door toeval hetzelfde

oordeel vellen gelijk aan het product van de proporties van de randtotalen. Voor het voorbeeld uit tabel 1 is dan de kans dat beide beoordelaars alleen op basis van toeval vinden dat een gebeurtenis plaatsvond (P_{pos}): .55 x .55 = .30. Het voorbeeld is in tabel 2 verder uitgewerkt.

Tabel 2. Proporties waarnemingen op basis van toeval bij de gegevens uit tabel 1.

		Observator A		
		Wel	Niet	Totaal
	Wel	.30	.25	.55
Observator B	Niet	.25	.20	.45
	Totaal	.55	.45	1

Volgens deze berekening zullen beide waarnemers op basis van toeval een overeenstemming van .50 bereiken (de som van P_{pos} en P_{neg}). Vullen we de gevonden getallen in de formule van kappa in dan resulteert 6 = (.80 – .50) / (1-.50) = .60.

Kappa kan ook voor meer beoordelaars worden berekend. Bij meer dan twee beoordelaars wordt van elk mogelijk paar beoordelaars P_o en P_e bepaald en wordt kappa berekend met de gemiddelden van deze waarden.

Kappa bereikt een waarde die varieert van maximaal +1 (perfecte overeenstemming tussen de beoordelaars), via 0 (de overeenstemming is net zo groot als de toevalskans) naar minimaal –1 (strikt tegenovergestelde beoordelingen van de beoordelaars).

Eigenaardigheden van Cohens kappa

In ons onderzoek observeerden studenten consulten aan de hand van 18 beoordelingsitems die geclusterd waren in 3 consultfases.[1] Van elk item beoordeelden zij of het wel of niet in het consult voorkwam. Aan het onderzoek namen 10 studenten deel die, onafhankelijk van elkaar, dezelfde vier op video opgenomen consulten beoordeelden. In tabel 3 worden de proportie overeenstemming en de waarden van kappa weergegeven voor de consultfases en voor de volledige consulten. Dit gebeurt zowel voor elk consult afzonderlijk als voor de vier consulten samen.

In de grijs gearceerde velden van tabel 3 doet zich het eigenaardige fenomeen voor dat de proportie overeenstemming hoog is (.87 tot .91) terwijl kappa bijna 0 of zelfs negatief is.

De achtergrond daarvan blijkt het gedrag van kappa te zijn bij goed beoordeelbare fenomenen die een scheve verdeling vertonen, c.q. als het fenomeen zeer frequent of juist heel weinig frequent voorkomt. Aan de hand van tabel 4 en 5 wordt gedemonstreerd wat er dan gebeurt.

In tabel 4 staan de resultaten van twee fictieve beoordelaars die een consultonderdeel observeren dat vaak aanwezig is

Tabel 3. Overeenstemming tussen studenten onderling bij de beoordeling van videoconsulten. Proportie overeenstemming (waarde van Cohens kappa).

	Consultnummer				
	1	2	3	4	1-4
Fase 1 (4 items)	.91 (.09)	.77 (.50)	.75 (.44)	.90 (-.03)	.83 (.47)
Fase 2 (9 items)	.75 (.27)	.80 (.50)	.77 (.55)	.63 (.42)	.74 (.46)
Fase 3 (5 item s)	.89 (.71)	.87 (.00)	.76 (.49)	.68 (.34)	.80 (.60)
Consult (18 items)	.82 (.44)	.81 (.55)	.76 (.51)	71 (.47)	.78 (.51)

en goed te beoordelen valt: P_{pos} = .80 en P_{neg} = .00. Hun totale proportie overeenstemming is .80.

Tabel 4. Proporties waarnemingen van twee fictieve observatoren bij een goed beoordeelbare en frequent voorkomende gebeurtenis.

		Observator A		
		Wel	Niet	Totaal
Observator B	Wel	.80	.10	.90
	Niet	.10	.00	.10
	Totaal	.90	.10	1

Als nu uit de randtotalen de kans op toevalsovereenstemming wordt berekend (tabel 5) dan valt die hoog uit: .82. Berekenen we vervolgens kappa, dan wordt die –0.11 ((.80-.82) / (1-.82)).

Tabel 5. Proporties waarnemingen op basis van toeval bij de gegevens uit tabel 4.

		Observator A		
		Wel	Niet	Totaal
Observator B	Wel	.81	.09	.90
	Niet	.09	.01	.10
	Totaal	.90	.10	1

Wisselen we in het voorbeeld P_{pos} en P_{neg} om, dan ontstaat een situatie waarin een consultonderdeel weinig voorkomt en de observatoren het ook daarover vaak met elkaar eens zijn. In die situatie veranderen de getallen voor P_e en kappa niet en blijft de paradox bestaan van een hoge proportie overeenstemming met een lage waarde van kappa.

Op basis van de waarde van kappa zou in beide gevallen ten onrechte geconcludeerd kunnen worden dat de beoordelaarsbetrouwbaarheid slecht is. Wat er werkelijk gebeurt, is dat onder de geschetste omstandigheden de schatting van de overeenstemming door toeval (op basis van de randtotalen) ten onrechte hoog uitvalt.

In ons onderzoek doet zich de hier beschreven situatie voor. Fase 1 van consult 4 (tabel 3) heeft bijvoorbeeld een proportie overeenstemming van .90 bij een kappa van -.03. Fase 1 bevat 4 items en de oorspronkelijke data laten zien dat bij twee daarvan onze studenten nooit van mening verschilden; bij de twee andere items heeft steeds slechts 1 van de 10 studenten een andere mening.

Bespreking

De kappacoëfficiënt blijkt een te negatief beeld te geven van de beoordelaarsbetrouwbaarheid in situaties waarbij het geobserveerde fenomeen scheef verdeeld is (het komt zeer frequent of juist weinig frequent voor) en bovendien goed te beoordelen is (waardoor de beoordelaars het in hoge mate met elkaar eens zijn).

De eerste dertig jaar nadat Cohen de kappacoëfficiënt beschreef is het rond deze eigenschap rustig gebleven. Pas sinds een artikel van Feinstein en Cicchetti uit 1990 is de discussie daarover op gang gekomen.[3] Sindsdien staat het bekend als de prevalentie-afhankelijkheid van kappa.[2-7] Het gaat om een probleem dat optreedt in elke situatie waarin de prevalentie van het fenomeen dat moet worden waargenomen sterk afwijkt van 50%.[8]

In het medisch onderwijs zal die situatie zich met name voordoen bij vaardigheidsexamens. Daarbij gebruiken we observatielijsten waarop items staan waarvoor we met ons onderwijs een prevalentie van 100% nastreven. Bovendien zorgen we ervoor dat die items goed te beoordelen zijn. Wordt dan de beoordelaarsbetrouwbaarheid met behulp van kappa gemeten, dan zal die ten onrechte laag uitpakken.

In dit tijdschrift lijkt dat fenomeen zich bijvoorbeeld voor te doen in het onderzoek van Boumans et al naar het effect van een observatorentraining bij een stationsexamen.[9] Ook in ons onderzoek observeerden studenten consultonderdelen waarvan het merendeel een hoge prevalentie had en goed te beoordelen was.[1] Achteraf gezien was Cohens kappa daarom geen goede maat om de beoordelaarsbetrouwbaarheid in uit te drukken.

In de aangehaalde literatuur wordt nog een tweede situatie genoemd waarbij de proportie overeenstemming en de waarde van kappa een geheel verschillend beeld van de beoordelaarsovereenstemming kunnen geven, namelijk als observatoren het structureel niet eens zijn over het al dan niet voorkomen van een fenomeen, bijvoorbeeld omdat het niet zo goed observeerbaar is. Dit wordt aangeduid met het bias-effect.[2 4-7]

Zowel van de prevalentie-afhankelijkheid als het bias-effect zijn de achtergronden inmiddels meer of minder theoretisch beschreven.[5-7] Verschillende auteurs hebben ten aanzien van hun invloeden ook suggesties voor correcties van kappa gedaan.[5-6] Recent is echter aannemelijk gemaakt dat daar beter geen gebruik van gemaakt kan worden.[8]

De prevalentie-afhankelijkheid en het bias-effect maken de interpretatie van kappa lastig. De prevalentie-afhankelijkheid zorgt er bovendien voor dat waarden van kappa uit verschillende studies niet zondermeer met elkaar vergeleken kunnen worden. Tussen twee onderzoeken zal immers de prevalentie van het fenomeen dat wordt geobserveerd zelden vergelijkbaar zijn.

De genoemde eigenschappen maken Cohens kappa minder geschikt als maat om de beoordelaarsbetrouwbaarheid in uit te drukken, zeker bij fenomenen met een hoge of lage prevalentie. Is er geen alternatief voor kappa, dan moet het advies van Cicchetti en Feinstein gevolgd worden om ook melding te maken van de proportie positieve en negatieve overeenstemming.[4]

Literatuur

1. Pols J, Andeweg ME, Boendermaker PM, Bosveld HEP, Terluin M. Hoe betrouwbaar en valide beoordelen studenten huisartsconsulten. Tijdschrift voor Medisch Onderwijs 2002;21(5):208-13.
2. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304(6840):1491-4.
3. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6):543-9.
4. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43(6):551-8.
5. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423-9.
6. Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. J Clin Epidemiol 1996;49(4):431-4.
7. Guggenmoos-Holzmann I. The meaning of kappa: probabilistic concepts of reliability and validity revisited. J Clin Epidemiol 1996;49(7):775-82.
8. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol 2000;53(5):499-503.
9. Boumans MTA, Scherpbier AJJA, Van Ooy A, Van der Vleuten CPM, Hoogenboom RJI, Schuwirth LWT. Het effect van een observatorentraining. Bulletin Medisch Onderwijs 1998;17(4):118-24.

De auteurs:

Dr. J. Pols is als arts-onderwijskundige werkzaam bij de afdeling Onderwijsontwikkeling en Kwaliteitszorg van het Onderwijsinstituut van de Faculteit der Medische Wetenschappen, Rijksuniversiteit Groningen / Stafgroep Organisatie en Ontwikkeling, Academisch Ziekenhuis Groningen.

Drs. H.E.P. Bosveld is methodoloog en werkzaam bij de disciplinegroep Huisartsgeneeskunde van de Faculteit der Medische Wetenschappen, Rijksuniversiteit Groningen.

Correspondentieadres:

Dr. J. Pols, Onderwijsinstituut, Faculteit der Medische Wetenschappen, Rijksuniversiteit Groningen, A. Deusinglaan 1, 9713 AV Groningen, tel. 050-363 7629, fax 050-363 3865, j.pols@med.rug.nl.

Summary

Introduction: *For the measurement of interobserver agreement Cohen's kappa (kappa) is often advised for nominal or ordinal observation scales, because it serves as a chance corrected measure of agreement.*

Problem: *We explored the reliability and validity of student observations of clinical consultations and sometimes found low values of kappa with high proportions of observer agreement (e.g., kappa= .03 with p= .91).*

Literature: *Kappa reaches too low values under circumstances where the phenomenon under observation occurs very frequently or very infrequently. In the literature this is known as the prevalence dependence of kappa. This happened in our study and it may be expected to occur regularly during skills examinations. Kappa also reaches too low values in situations where observers often disagree on the (non-)occurrence of a phenomenon. This is known as the bias-effect.*

Conclusion: *Kappa is less suitable to measure observer agreement if the phenomenon under observation is not evenly distributed. If kappa is the only alternative, the proportions of positive and negative agreement should also be mentioned. (Pols J, Bosveld HEP. (Refrain from) measuring observer agreement with Cohen's kappa? Dutch Journal of Medical Education 2003;22(5):229-234.)*